DOCUMENT RESUME

ED 421 525 TM 028 850

AUTHOR Davey, Tim; Parshall, Cynthia G.
TITLE New Algorithms for Item Selection and Exposure Control with Computerized Adaptive Testing.
PUB DATE 1995-04-00
NOTE 25p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 18-22, 1995).
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Ability; *Adaptive Testing; *Algorithms; *Computer Assisted Testing; *Efficiency; *Selection; Simulation; Test Construction; *Test Items; Testing Problems
IDENTIFIERS *Item Exposure (Tests); Test Security

ABSTRACT

Although computerized adaptive tests acquire their efficiency by successively selecting items that provide optimal measurement at each examinee's estimated level of ability, operational testing programs will typically consider additional factors in item selection. In practice, items are generally selected with regard to at least three, often conflicting, goals: (1) to maximize test efficiency by measuring examinees as quickly and accurately as possible; (2) to protect the security of the item pool by controlling the rates at which popular items can be administered; and (3) to assure that the test measures the same composite of multiple traits for each examinee by balancing the rates at which items with different content properties are administered. This paper focuses on the goals of maximizing test efficiency and controlling item exposure rates, avoiding a discussion of content balance. Problems in existing algorithms for accomplishing these goals are outlined and illustrated, and some alternative algorithms that offer at least a partial solution are presented. Posterior weighted information is suggested as a new item selection method, and its usefulness is demonstrated through a simulation. Conditional exposure control is suggested to control exposure rate, and a similar simulation is presented to demonstrate its usefulness. (Contains one table, seven figures, and seven references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 421 525

New Algorithms for Item Selection and Exposure Control with Computerized Adaptive Testing

Tim Davey

Cynthia G. Parshall

American College Testing

Abstract

Computerized adaptive testing (CAT) offers the prospect of both reducing testing time and increasing measurement precision when compared to conventional pencil-and-paper tests. Although adaptive tests acquire their efficiency by successively selecting items that provide optimal measurement at each examinee's estimated level of ability, operational testing programs will typically consider additional factors in item selection. In practice, items are generally selected with regard to at least three, often conflicting goals: 1) to maximize test efficiency by measuring examinees as quickly and as accurately as possible, 2) to protect the security of the item pool by controlling the rates at which popular items can be administered, and 3) to assure that the test measures the same composite of multiple traits for each examinee by balancing the rates at which items with different content properties are administered.

This paper focuses on the goals of maximizing test efficiency and controlling item exposure rates, avoiding discussion of content balance. While a number of algorithms for accomplishing these goals have been developed, all are problematic to some extent. We briefly sketch the nature of these problems, and then present alternative algorithms that offer at least a partial solution.

U.S. DEPARTMENT OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Cynthia Parshall

TM028850

Paper presented at the annual meeting of the American Educational Research Association, April 18-22, 1995, San Francisco

New Algorithms for Item Selection and Exposure Control with Computerized Adaptive Testing

1. Introduction

Computerized adaptive testing (CAT) offers the prospect of both reducing testing time and increasing measurement precision when compared to conventional pencil-and-paper tests. Although adaptive tests acquire their efficiency by successively selecting items that provide optimal measurement at each examinee's estimated level of ability, operational testing programs will typically consider additional factors in item selection. In practice, items are generally selected with regard to at least three, often conflicting goals: 1) to maximize test efficiency by measuring examinees as quickly and as accurately as possible, 2) to protect the security of the item pool by controlling the rates at which popular items can be administered, and 3) to assure that the test measures the same composite of multiple traits for each examinee by balancing the rates at which items with different content properties are administered.

This paper focuses on the goals of maximizing test efficiency and controlling item exposure rates, avoiding discussion of content balance altogether. While a number of algorithms for accomplishing these goals have been developed all are problematic to some extent. We briefly sketch the nature of these problems, and then present alternative algorithms that offer at least a partial solution.

2. Maximizing Test Efficiency

The most basic principle of adaptive testing is to select items that measure optimally in some sense at the examinee's level of ability. However, since the purpose of testing is to ascertain this level, new items are actually chosen with respect to *provisional* estimates of ability that are based on those items already administered. Provisional estimates, either maximum likelihood or bayesian, are revised following each new response and are subject to considerable error, especially early in the test. Some efficiency is then necessarily lost when items target an estimated ability that lies some distance from the examinee's true level.

Some Current Procedures

The common criteria for item selection differ in how they use provisional results to determine the next, best item for administration. The *maximum information* (MI) criterion (Brown & Weiss, 1977), selects the item that has the largest Fisher information value at the examinee's current ability estimate. A second common criterion, *maximum posterior precision* (MPP) (Owen, 1975), is based on successively reducing the variance of the posterior ability distribution. More recently, Veerkamp and Berger (1994) developed a general framework of weighted information-based item selection criteria into which MI fits neatly as a special case. Several other selection criteria are encompassed by this framework, including the new criterion suggested here.

Maximum information. Maximum information item selection is computationally very simple during testing because the heavier calculation of information functions can be done up front and their results tabled on a discrete grid of ability values. A further step usually taken is to rank order items by the amount of information that they provide at each of the gridded ability values and list the items in this order in the columns of an *information table*. Item selection is then as easy as finding the column of the information table that brackets the provisional ability estimate, and pulling the top rated

unpresented item in that column off the stack. Unfortunately, error in the provisional estimates will often lead to items being selected from columns that do not cover the true ability value. This problem is exacerbated when items are highly discriminating, as those near the top of the information table columns invariably are. Such items are generally fairly tightly focused, discriminating well over a narrow ability range and considerably less well outside this range. An item measuring well at the provisional estimate may therefore measure poorly at the true ability value.

Maximum posterior precision. The MPP selection procedure recognizes that provisional estimates contain error. Items are therefore selected based on the entire posterior distribution of ability rather than a single point estimate. The item selected may not be the most informative at the particular provisional ability level, or at any other ability level for that matter. The selected item is instead a compromise that measures well on average across the high density region of the posterior distribution. This approach is more conservative in nature and will often yield superior results. However, the MPP approach cannot be based on information tables and is therefore much more computationally intensive than MI. The procedure must continually search the entire bank of unpresented items to find the one that leads to maximal reduction of the posterior variance, a process that can be extremely time consuming for even moderately large item banks.

Weighted information criteria. Veerkamp and Berger (1994) developed a general framework for item selection under which weights of various sorts are applied to the columns of the information table. During item selection, the information values provided by each item at each ability level are multiplied by these weights and summed. The item with the largest weighted sum information is then chosen for administration. The maximum information criterion can be viewed as a special case under which all weight is massed on the single column of the table that contains the provisional ability estimate. This led Veerkamp and Berger to term MI the *point information criterion*. They also considered an *interval information criterion*, which equally weights some number of table columns that surround and include the column containing the provisional ability estimate. This was extended further to the *likelihood weighted criterion*, which applied weights to each column that were proportional to the likelihood of the examinee's true ability falling into that column. While simulation results suggest that the likelihood weighted criteria outperforms maximum information, particularly with short tests, one drawback to the approach is the extensive computations required during testing to construct the likelihood functions.

A New Procedure -- Posterior Weighted Information

The new item selection procedure retains many of the attractive features of some current methods, while avoiding certain of their drawbacks. The new procedure is like the MPP and the likelihood weighted criterion in acknowledging that provisional ability estimates are subject to error. It is in fact quite similar in nature to the likelihood weighted criterion and falls into the general class of weighted information methods. However, the procedure is also computationally simple, as its weights are based on Owen's (1969, 1975) quick approximation to the posterior ability distribution rather than the more computationally intensive likelihood function. A second advantage to weighting by the posterior distribution rather than the likelihood function, is that the latter is somewhat unstable and is almost certain to be unbounded early in the test. The posterior distribution shares neither of these attributes.

For maximum efficiency, the *posterior weighted information* procedure is based on an information table whose columns correspond to ability values or ranges. In each column, item numbers and associated Fisher information values are listed in order of descending information. This table is used to select items as follows:

1. Compute Owen's (1969) approximation to the posterior distribution following each item response. This approximation is a normal distribution with specified mean and variance.

2. Use the approximate posterior distribution to assign weights to each column of the information table that are proportional to the column including the examinee's true ability.

3. Identify the k top-ranked items in each column of the table, where 10 would be a reasonable value for k. Form the union of the different sets of items identified across columns.

4. Compute the weighted sums of information values across columns for each item in the aggregate set.

5. Select the item with the largest weighted information sum for administration.

Evaluation

The posterior weighted information (PWI) procedure was compared to maximum information in a simulation study. While simulation studies are not generally desirable, there is really no effective alternative to them in evaluating adaptive testing procedures. The problem with such studies is that it is all but impossible to design and conduct a series of simulations extensive enough to assess the procedures with complete generality. Therefore, no attempt was made to do so. The study reported here is quite modest in scope and should therefore be viewed more as a starting point than as a final answer.

All simulations were conducted using two item pools, the first containing 100 items, the second 200, with the second pool subsuming the first. Both pools were designed to approximate reality by being based on item parameters drawn from a currently operational adaptive testing program. Simulated examinees were drawn from a standard normal population and administered a fixed length CAT of thirty items. Items were selected by either the PWI or MI criteria, with provisional ability estimates computed by Owen's bayes in either case. Simulation results are based on very large (100,000) examinee samples so that their stability should not be an issue.

Method. Every item pool contains some fixed amount of information at each ability level, a quantity computed by summing the item information functions across the items in the pool. The 100 and 200 item total pool information functions computed in this way are shown as Figure 1. These functions fix the upper bound of how precise any adaptive test can be at each ability level. The information functions of each selected adaptive test can be computed in the same way, by summing item information functions across the selected items. Since an effective selection procedure quickly identifies and administers items that are most informative at the true ability of the examinee, the quality of the competing procedures can be compared by determining how efficiently each method extracts the information that the item pool has available at the true ability level. This was done by evaluating both the test and the pool information functions at the true ability level of each examinee and comparing their values. This process went as follows:

1. Sample an ability value from a standard normal distribution. Denote this "true" ability by θ_t, and use it to generate responses to the thirty-item adaptive tests chosen by each selection procedure.

2. For each item chosen, compute $I_j(\theta_t)$, the information that item j provides at the true ability, θ_t. Also compute the total pool information at θ_t, the limiting value of any selected test's precision. Denote this value by $I_{max}(\theta_t)$

3. Compute the cumulative information through each selected item, one through thirty. Denote these by $I^n(\theta_t)$, where

$$I^n(\theta_t) = \sum_{j=1}^{n} I_j(\theta_t) \quad n = 1,2,\ldots,30$$

Each of these values shows the amount of information accumulated through a given number of items, making it possible to track the rate at which the test information approached the limiting pool information.

4. Divide each cumulative information value by the total pool information, $I_{max}(\theta_t)$. The result is the proportion of total information in the pool that the item selection procedure was able to extract with tests of varying lengths.

5. Steps one through four were repeated 100,000 times for each of the two item selection procedures and the results accumulated and averaged.

Results

The above results are presented in three different ways:

1. **By ability level.** Figure 2 presents a series of plots displaying the proportions of pool information extracted with each test length at nine levels of true ability. Each plot shows four curves, one for each combination of pool size (100 and 200 items) and item selection procedure (Old=maximum information, New=posterior weighted information). The higher pair of curves on each plot show the results for the 100 item pool, while the lower pair pertains to the 200 item pool. Each curve traces the proportions of information extracted for tests of one through thirty items. The first plot in the series shows the results for those cases where the true ability equaled -4. Remaining plots are for true abilities of -2.4, -1.6, -.8, 1.6, 2.4, and 4, respectively. Some observations:

   - Selection under PWI is generally superior to MI, particularly for extreme abilities. Both selection procedures begin by administering items of moderate difficulty, which prove an excellent choice when testing middle-ability examinees. Since the improvement offered by PWI is expected to be realized early in the test, little effect is noticed when both procedures start well. However, when true abilities are extreme, the MI procedure seems to take longer to shift its focus from administering middle-difficulty items, giving the advantage to PWI for these examinees.

   - The improvement offered by PWI is more noticeable with the smaller item pool than it is with the larger. This was not expected. It was thought that MI would make poorer

decisions when presented with larger numbers of highly discriminating items. This effect requires further investigation.

2. **By test length.** Figure 3 contains a series of plots presenting the same information as above in a different way. Plots in this series show results across ability levels for a test of a given length. Each plot again contains two pairs of curves, with the upper and lower pairs relating to the 100 and 200 item pools, respectively. Each curve traces across ability levels the relative information, or the proportion of total pool extracted, for both item selection methods. The first plot in the series shows results for a five-item test, with subsequent plots covering 10, 15, 20, 25, and 30 items.

   - All curves have significant dips for central ability values, not because less information is being extracted here but because so much more is available. In fact, the raw (rather than relative) information plots are peaked in the center, because both procedures start with items that discriminate well in the center and because more highly discriminating items are available there.

   - The superiority of PWI in measuring extreme abilities is emphasized, as is the lack of significant difference between the methods with central abilities.

3. **Marginal reliabilities.** Figure 4 summarizes results in still another way, by plotting the marginal reliabilities of tests of various lengths. The marginal reliability is the CAT analogue to the conventional measure of test reliability, and is computed by averaging error variances across the ability scale relative to some target distribution of examinee ability (Thissen and Mislevy, 1990). Here, the ability distribution was assumed to be standard normal. Each curve on these plots traces the increase in test reliability with increasing test length. More effective item selection is indicated by more rapidly increasing trace lines.

   - PWI was again shown superior, with this superiority becoming evident somewhere between the third and seventh test item. Both procedures struggle equally early in the test as they try to recover from having inappropriately selected several middle-range difficulty items for examinees who are not of middle-range ability.

   - The early advantage shown by PWI is soon erased by the law of diminishing returns. This is not at all surprising, as by the seventh or eighth test item both selection procedures are largely in agreement as to which item to present next.

Conclusion

Item selection by the PWI criterion does offer advantages over the more straightforward maximum information. However, these advantages seem limited to short tests and extreme examinees, suggesting a hybrid approach in whichn PWI would be used early in the test before switching to MI later. How PWI would fare against other weighted information and the MPP is a matter for future inquiry.

3. Controlling Item Exposure Rates

A common finding with all optimal approaches to item selection is that certain items will be administered to nearly every examinee. Furthermore, a small number of the available items will account for a large proportion of item administrations. This situation is best avoided for reasons of test security. Accordingly, several procedures for controlling item exposure rates have been devised.

Some Current Procedures

The simplest of the current exposure rate control procedures is sometimes called the 4-3-2-1 algorithm (McBride & Martin, 1983). This procedure asks an item selection rule to choose not only the best item for administration at a given point, but the second, third and fourth best items as well. The best item is then actually presented only 40% of the time. The second, third and fourth best items are presented 30%, 20% and 10% of the time, respectively. Items selected but not presented are returned to the bank for later use. Because items can be repeatedly selected, high ranking items are usually eventually administered. Therefore, while the procedure serves to shuffle the order in which items are presented it does not actually control item exposure rates.

A more elaborate procedure assigns each item an *exposure parameter*, which takes on a value between zero and one (Sympson & Hetter, 1985). Items selected as best according to rules like MI or PWI are actually administered with the probability given by their exposure parameter. Items selected but not administered are set aside and not reselected unless the item pool becomes exhausted.

Exposure parameters are estimated by repeatedly simulating the administration of several thousand adaptive tests. Following each simulation, the frequency with which each item was presented is tallied and compared to some subjective standard. For example, it may have been decided that no item should be administered to more than 25% of all examinees. The exposure parameters for items whose frequency of use exceeds this standard are then successively adjusted downward as the cycle of simulations continues. The cycle ends when the exposure parameters have stabilized and no items exceed the usage standard.

The simulations used to set exposure parameters are conducted with respect to some target or projected distribution of examinee ability. Items that discriminate well near the center of this distribution are subject to overuse and will therefore be assigned fairly small exposure parameters. Conversely, items that discriminate well in the tails of the target distribution will be rarely used in any case, so that most of their exposure parameters would be set near or equal to one. Item exposure rates may therefore actually be controlled only to the extent that the observed ability distribution is similar to that projected.

A New Procedure -- Conditional Exposure Control

The true goal of an exposure control procedure is not strictly to limit the frequency of item use, but rather the overlap across tests. Although a given procedure may serve to control exposure rates, it does not necessarily minimize the extent to which tests overlap across examinees. This is because exposure probabilities are treated unconditionally: the probability of administering a selected item is constant and independent of those items that have been previously administered or are likely to be subsequently administered. However, adaptive tests will often result in clusters or sets of items which appear together with unwelcome frequency. This occurs even when the Sympson-Hetter procedure is

used. One solution is make item exposure probabilities conditional on those items that have already appeared.

The proposed procedure is very similar to Sympson and Hetter, with one important difference. Sympson and Hetter assign each item a single, global exposure parameter. The new procedure provides a probability for exposure which is conditioned on those items previously presented to an examinee. Thus, if two or more items tend to appear together exceptionally often, the new exposure control procedure will limit the probability of the remaining items in this set appearing once any item in the group has been administered.

The procedure is based on an $n \times n$ table of exposure parameters, where n is the number of items in the pool. Diagonal elements of this table contain unconditional exposure parameters similar to those used under Sympson-Hetter. The off diagonal entries are the conditional parameters that control the frequency with which pairs or cluster of items occur. The Sympson-Hetter procedure then can be viewed as a special case of the proposed procedure, in which all of the off-diagonal values are set to unity.

The procedure can be illustrated using the following small example of an exposure table.

		Item				
		1	2	3	4	5
Item	1	.8	.8	.2	1.	.4
	2	.8	1.	1.	1.	1.
	3	.2	1.	.4	.8	.6
	4	1.	1.	.8	1.	1.
	5	.4	1.	.6	1.	.7

This table is symmetric, with values above the diagonal equal to those below. The diagonal elements give evidence of an item's popularity, with values at or near one indicating that the item is not selected frequently enough to warrant protecting it against overadministration. Item three, on the other hand, has a diagonal entry of .4, suggesting that it is quite attractive and will therefore be allowed to be administered only a small number of the occasions it is selected. The off-diagonal entries are interpreted in much the same way, except the values now pertain to pairs rather than to individual items. Low off-diagonal values are used to prevent popular pairs or sets of item from co-occurring. The item one-three combination is seen to be especially troublesome and will therefore be strictly controlled. However, item four is apparently not closely associated with other items in the table and so can freely appear with any.

The table is used during the test to determine conditional exposure parameters through the following procedure:

1. Select an item using PWI, MI or some other selection algorithm.

2. Enter the exposure table and find the diagonal entry for the selected item. Denote this value by e_{ii}. Also find the off-diagonal entries involving both the selected item and those items administered previously in the test. Denote these values by e_{ij}, where j ranges from one to the number of items so far administered. (Returning to the sample table, suppose that item five

has been selected and that items two and three have already been administered. Then $e_{55} = .7$, and $e_{52} = 1.$, $e_{53} = 6$).

3. Compute the conditional probability of administering the selected item by taking the mean of the set of e_{ij} values and multiplying the result by e_{ii}. Administer the selected item with this probability. (In the example, the probability of administering item five give that items two and three have appeared already is: .7 (1 + .6)/2 = .56). Methods for characterizing the distribution of the e_{ij} other than by its average can be considered but were not explored extensively here.

This procedure was expected to lead to a more balanced use of the item pool and to reduce the overlap in tests both for repeat test-takers (or examinees with similar ability) and across peers (examinees of differing ability).

Estimating the exposure table

Before the above procedures can be implemented, the entries in the exposure table must be determined. This is done through a series of simulations, much as is the case with Sympson-Hetter. This process runs as follows:

1. Set upper limits on the frequency with which individual items and item pairs are allowed to appear. These limits are somewhat subjective and are a function of pool size, average test length, and desired level of security. Limits for individual items are expressed as the proportion of tests in which an item is allowed to appear. For example, items may be restricted to occurring in at most 15% of the tests administered. Pairwise limits are expressed differently, as described below.

2. Initialize both the diagonal and off-diagonal cells of the exposure table to unity.

3. Simulate a large number (5000+) of adaptive tests, drawing ability parameters from some target distribution. Keep track of the number of times each item and pair of items appears.

4. Decrease the diagonal entry of the exposure table for any item that appears more frequently than the limit allows. We suggest this be done by multiplying the current value by .95. Allow diagonal entries for items that occur less frequently than is allowed to increase (to a maximum of 1) by multiplying by 1.04. This is done because any item may appear too frequently in a single simulation and should not have its parameter reduced due to this chance occurrence.

5. Consider for each pair of items the following contingency table, which is constructed following each set of 5000 simulations:

		Item i		
		Appeared	Did not Appear	Total
	Appeared	A	B	A+B
Item j	Did not Appear	C	D	C+D
	Total	A+C	B+D	A+B+C+D

Of concern is cell A, the number of tests in which both items i and j appeared. Ideally, this number should be no larger than would be expected by chance, indicating no positive association between the pair of items. While this hypothesis can be tested by chi-square statistics, these must be modified to ignore negative associations, which are in some sense desirable. Here, the statistic used was:

$$\chi^2 = \begin{cases} \dfrac{(A+B+C+D)(A \cdot D - B \cdot D)^2}{(A+C)(B+D)(C+D)(A+B)} & \text{if } (A \cdot D - B \cdot D) \geq 0 \\ 0 & \text{if } (A \cdot D - B \cdot D) < 0 \end{cases}$$

If these statistics exceeded a specified maximum value, the i,j entry of the exposure table is decreased by being multiplied by .95. As with the diagonal values, entries were increased for those item pairs whose statistic fell below the maximum value in order to avoid capitalizing on chance. The problem of multiple comparisons certainly must be considered when specifying a maximum value for the chi-square statistics. We used values of 7.5 for the 100 item pool and 10 for the 200 item pool. However, some experimentation would be required before values appropriate for other item pools, other test lengths, other ability distributions, and other security goals could be determined.

6. Repeat steps 3 through 5 until the exposure table stabililizes, usually after 200 - 300 iterations.

Evaluation

The conditional exposure control procedure was evaluated through a simulation study similar to that used with the item selection procedures. It was compared to unconditional control (Sympson-Hetter) and no control at all using the same two item pools as before, the first with 100 items, the second with 200. However, variable rather than fixed-length tests were administered. Testing ended when the posterior variance feel below .12, or when a maximum of fifteen items had been administered. The great majority of the tests administered (80%) met the posterior variance stopping rule rather than administering the full fifteen items. All simulations used huge numbers (50,000) of examinees drawn from standard normal distributions.

Two types of test overlap were examined: test-retest and peer-to-peer. The first case involves examinees retesting without any intervening treatment to change their ability level. Test overlap is expected to be high in this case because items will be drawn from the same part of the item pool during both tests. Peer-to-peer overlap looks at the communality of tests given to examinees of randomly dissimilar abilities. This is important in determining how much "help" previously tested examinees could provide to their colleagues who are about to test.

When comparing the conditional and unconditional methods of exposure control, it is critical to match the unconditional levels of item exposure. For example, if the unconditional procedure limited items to appearing on at most 25% of the tests administered, the conditional procedure must use this same limit. This is important because test overlap can be driven down simply by reducing the frequency with which items can appear. In the study reported here, unconditional appearance limits of 25% and 15% were imposed for use with the 100 and 200 item pool, respectively.

The basic results are summarized in Table 1. Using no exposure control obviously invites disaster, with mean overlap rates of 43.9% and 71.3% with the 100 item pool. These rates drop only slightly with the doubled pool size, proof that large pools are not in themselves sufficient to insure security. Conditional control of item exposure was found uniformly superior to unconditional control. As might be expected, the degree of improvement is greater in the test-retest case, where item clusters are more likely to form. Because tests were of variable length, reliability was constant across exposure control method and pool size. However, the numbers of items needed to achieve this level of reliability did differ somewhat. Both methods of exposure control lead to similar decreases in test efficiency, an effect of about one item with the smaller pool and 2.5 items with the larger. This decrease in relative efficiency of the exposure controlled tests with the larger pool is likely the result of stricter control (15% versus 25% maximum appearance rates) rather than a general finding. The no-control tests are substantially more efficient with better-targeting 200 item pool than they are with the less adaptable smaller pool.

The above results are emphasized by the plots in Figure 5, which display the full distributions of overlap rates across the 50,000 tests administered in each condition. The distribution for the conditional procedure is consistently shifted to the left of that for the unconditional and no-control procedures, confirming lower rates of test overlap.

Figure 6 plots the exposure rates of each item under the three types of control. These figures, much more readily interpreted in color, show that both the conditional and unconditional procedures are successful in limiting item exposure rates to 25% and 15% in the two pools. Without control, a handful of items are seen to appear very frequently while the remainder of the pool goes largely unused. A careful eye reveals that conditional control forces a somewhat more balanced use of the pool than does unconditional.

Conclusion

Conditional control of item exposure rates is successful in reducing the extent of item overlap across tests administered. Furthermore, the degree of reduction is greatest for what may be the most serious challenge to test security, that of examinees who test repeatedly in a short period of time. While any type of exposure control will produce a decrease in test efficiency, conditional control does not seem to impose any costs above that observed with more conventional unconditional methods.

References

Brown, J.M. & Weiss, D.J. (1977). *An adaptive testing strategy for achievement test batteries* (Research Report 77-6). Minneapolis: University of Minnesota, Psychometric methods Program.

McBride, J.R. & Martin, J.T. (1983). Reliability and validity of adaptive ability tests in a military setting. In D.J. Weiss (Ed.), *New horizons in testing* (pp.223-226). New York, Academic Press.

Owen, R.J. (1969). *A Bayesian approach to tailored testing* (Research Report 69-92). Princeton, NJ: Educational Testing Service.

Owen, R.J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association*, 70, 351-356.

Sympson, J.B. & Hetter, R.D. (1985). Controlling item-exposure rates in computerized adaptive testing. *Proceedings of the 27th annual meeting of the Military Testing Association* (pp. 973-977). San Diego, CA: Navy Personnel Research and Development Center.

Thissen, D. & Mislevy, R.J. (1990). Testing Algorithms. In H. Wainer, *Computerized Adaptive Testing: A Primer* (pp. 103-136). NJ: Lawrence Erlbaum Associates.

Veerkamp, W. J. J. & Berger, M. P. F. (1994). *Some new item selection criteria for adaptive testing* (Research Report 94-6). Enschede: University of Twente, Faculty of Educational Science and Technology.

Table 1

Item Exposure Results

	100 Item Pool			200 Item Pool		
	None	Unconditional	Conditional	None	Unconditional	Conditional
Peer-to-Peer Overlap	43.9	22.7	18.9	40.6	14.4	12.0
Test-Retest Overlap	71.3	40.7	27.0	66.2	28.4	19.3
Average Reliability	.91	.90	.90	.91	.91	.90
Average Test Length	12.8	13.8	13.9	11.5	14.0	14.1

Figure 1

Pool Information Functions

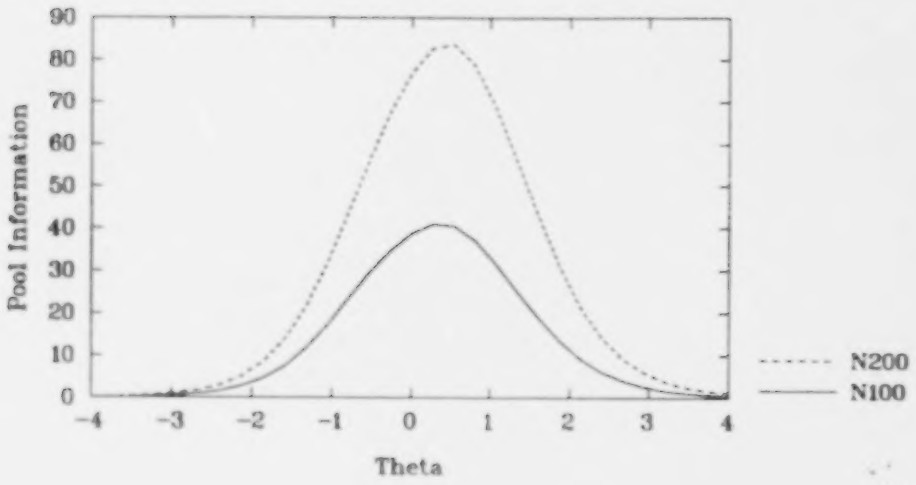


Figure 2

Results By Ability Levels

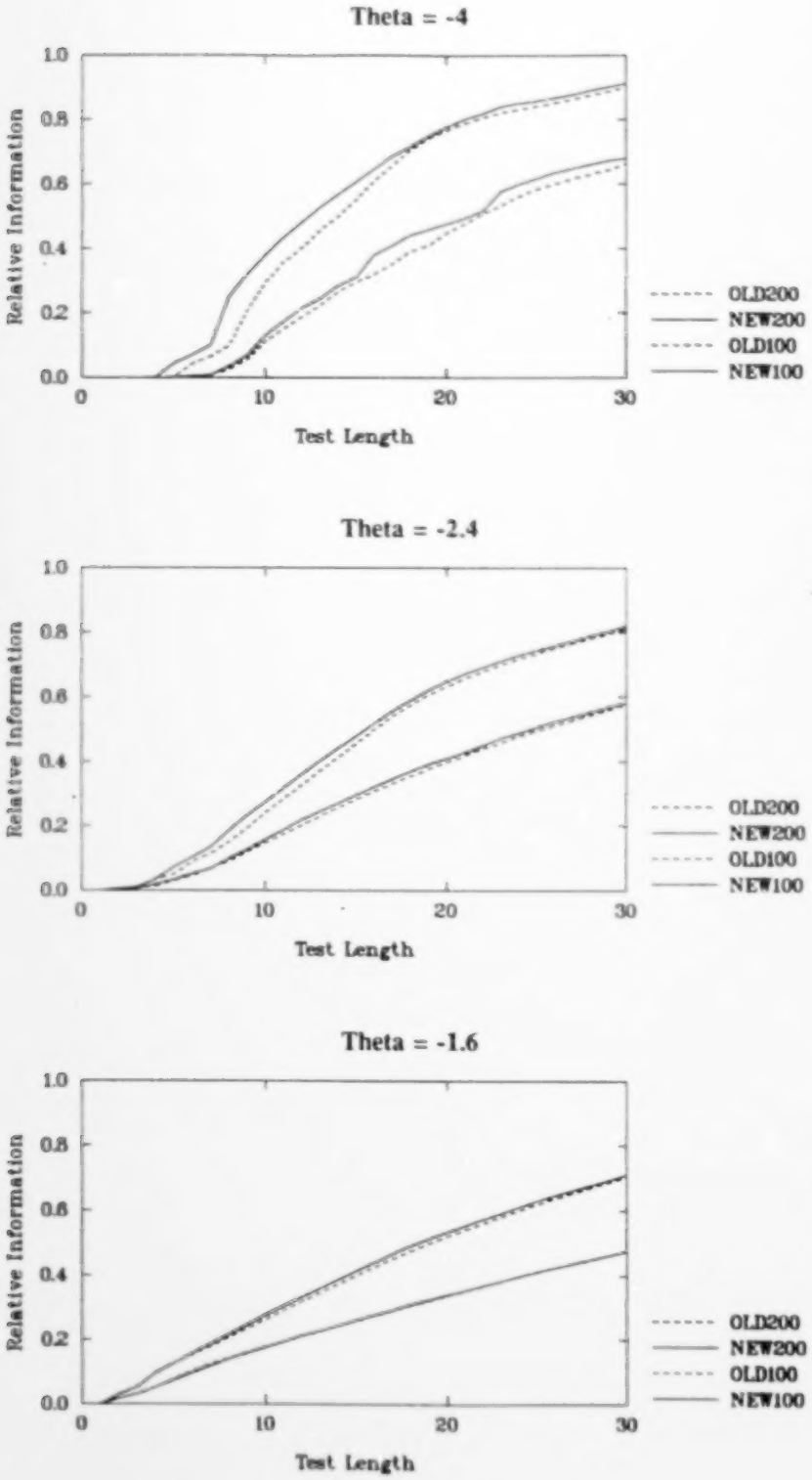


Figure 2 (cont.)

Results By Ability Levels

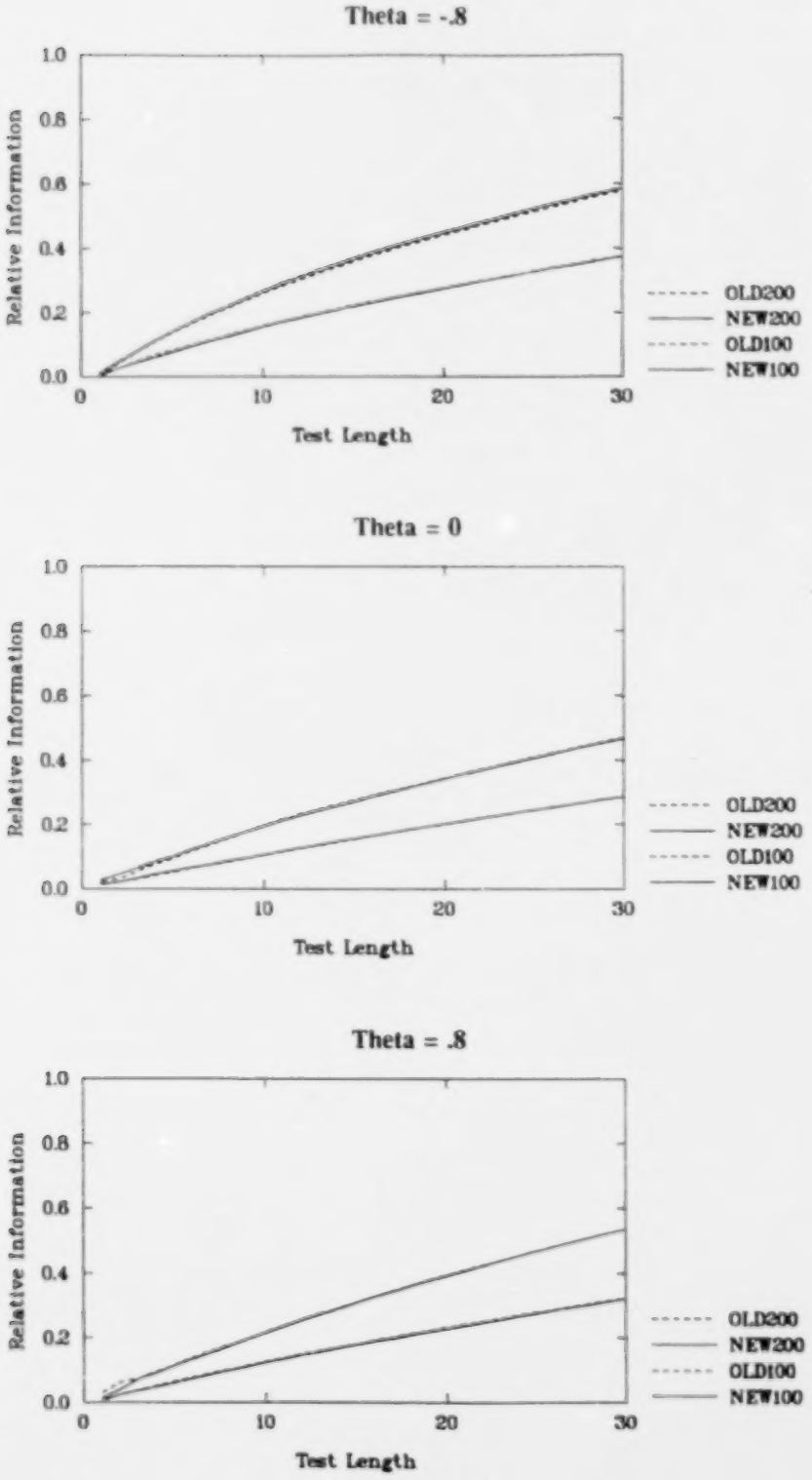
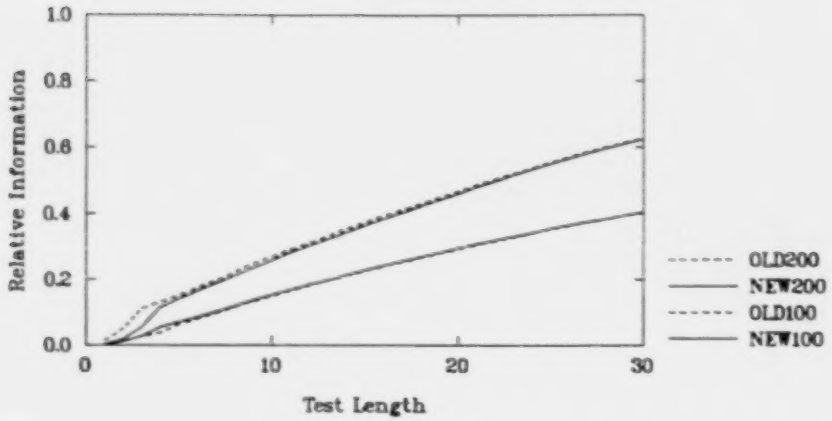


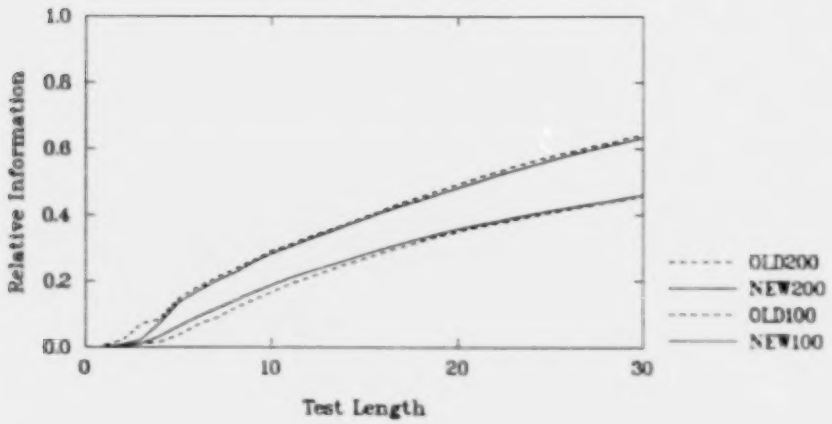
Figure 2 (cont.)

Results By Ability Levels

Theta = 1.6



Theta = 2.4



Theta = 4

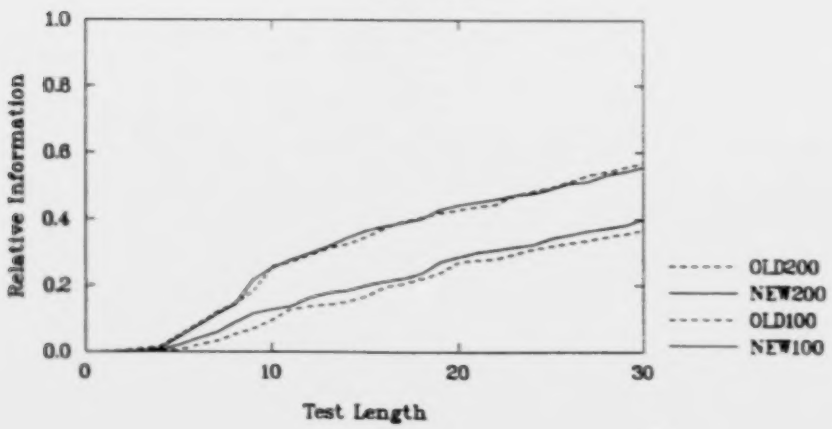


Figure 3

Average Relative Information

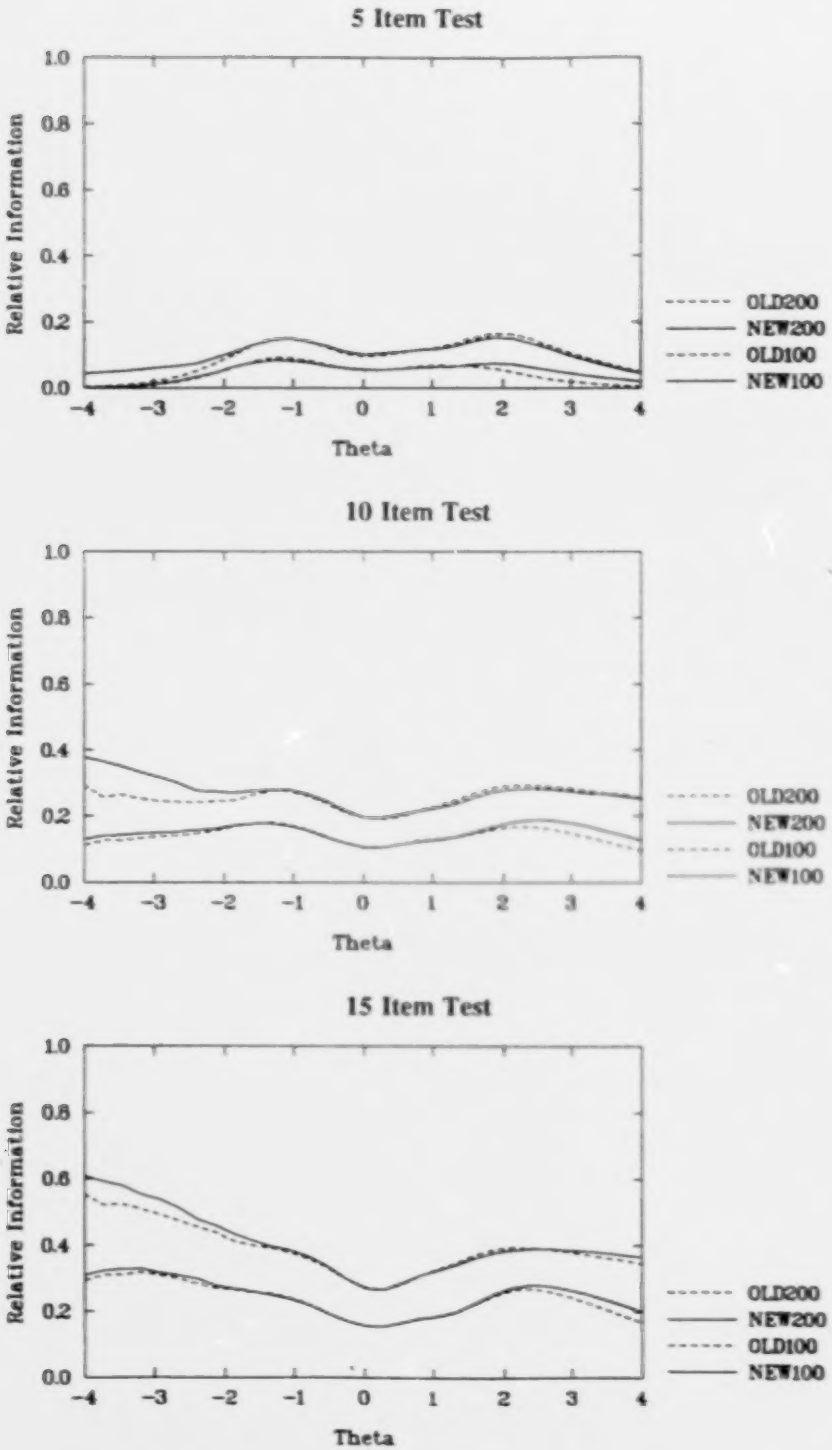


Figure 3 (cont.)

Average Relative Information

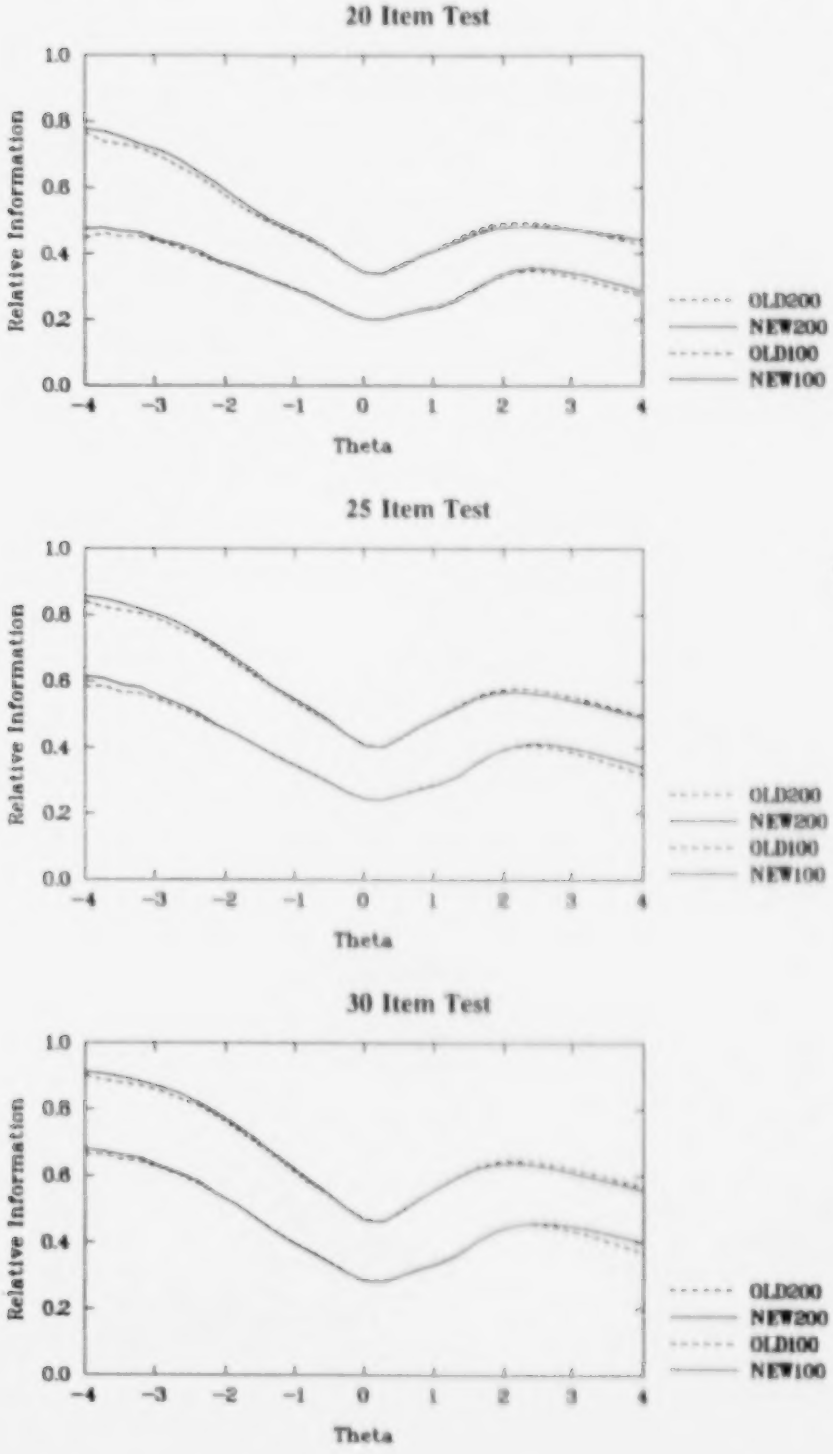


Figure 4

Marginal Test Reliabilities

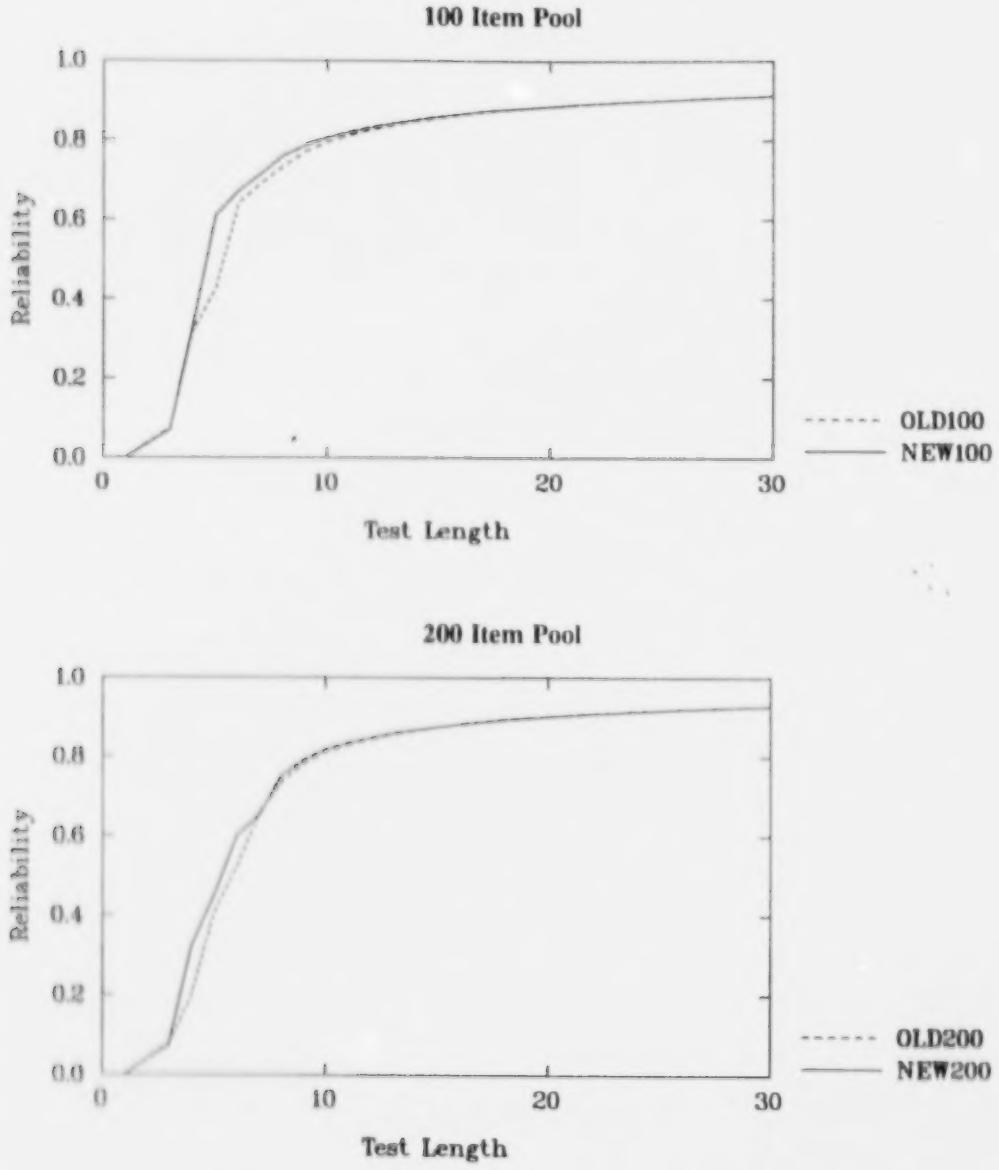
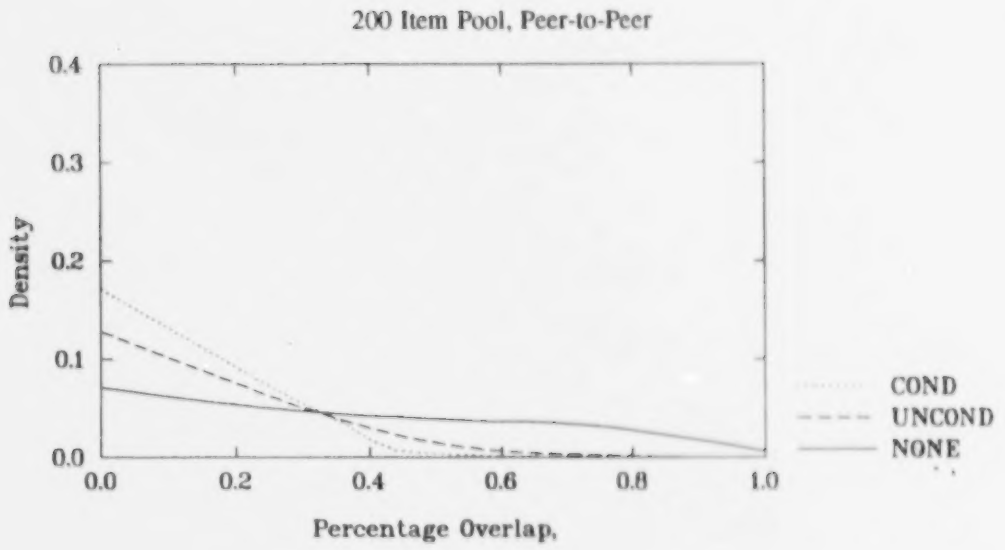


Figure 5

Test Overlap Rates



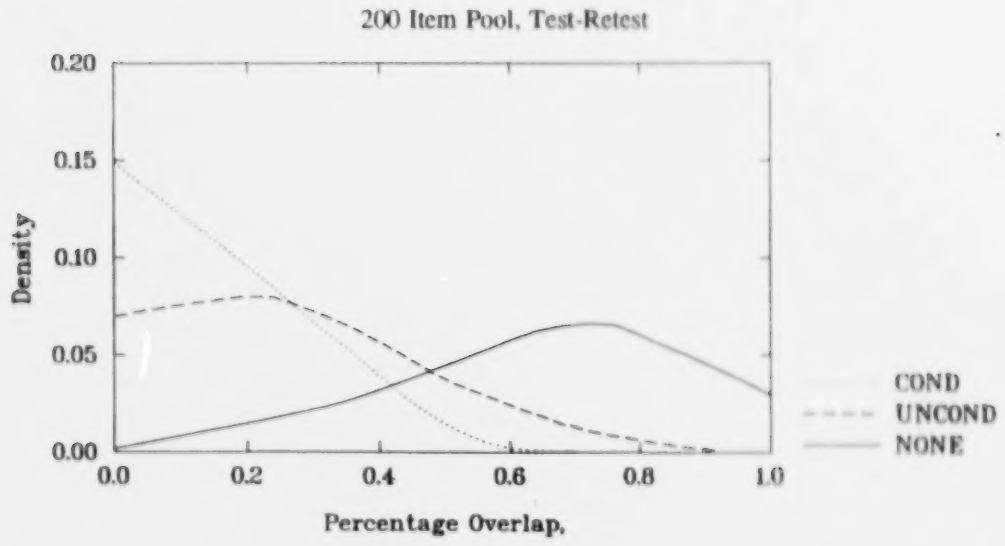
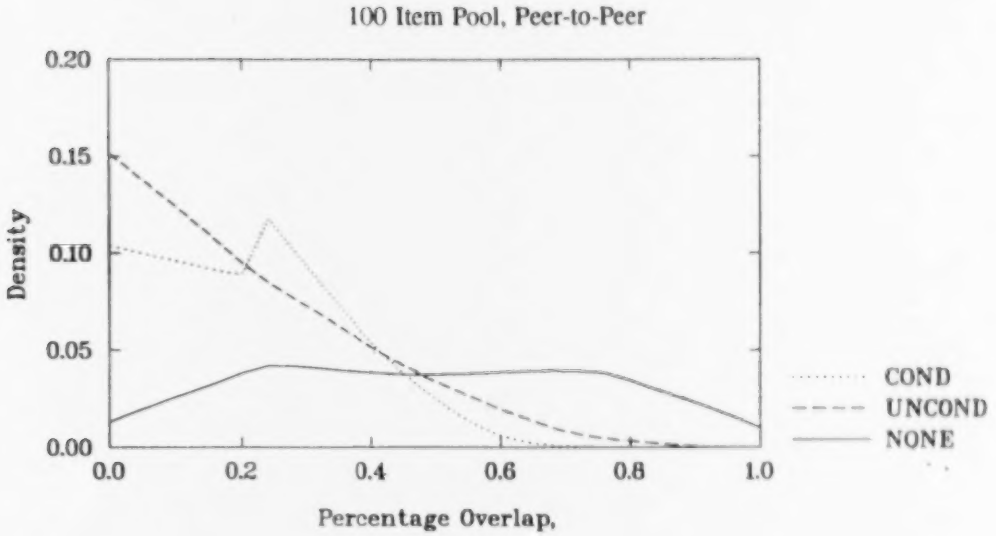


Figure 5 (cont.)

Test Overlap Rates



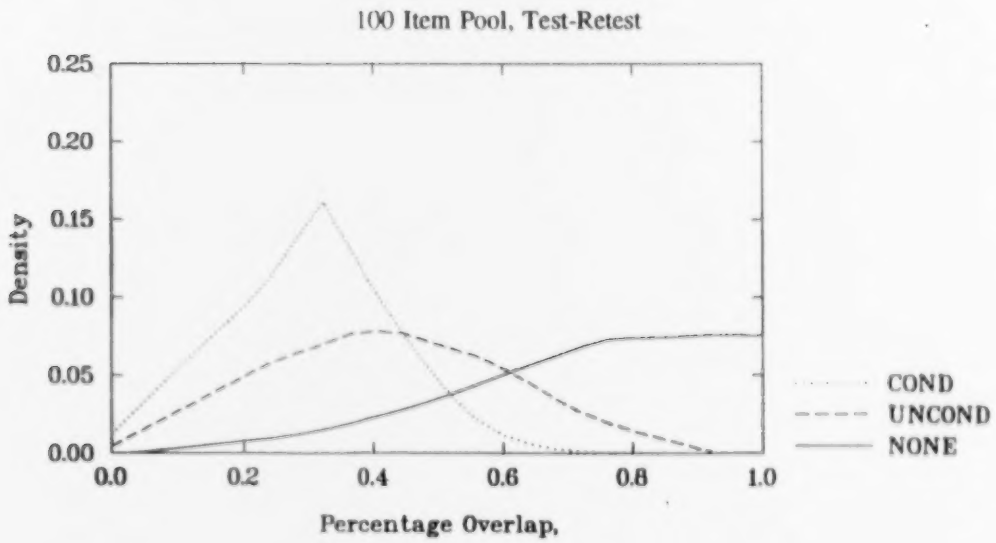


Figure 6

Item Exposure Rates

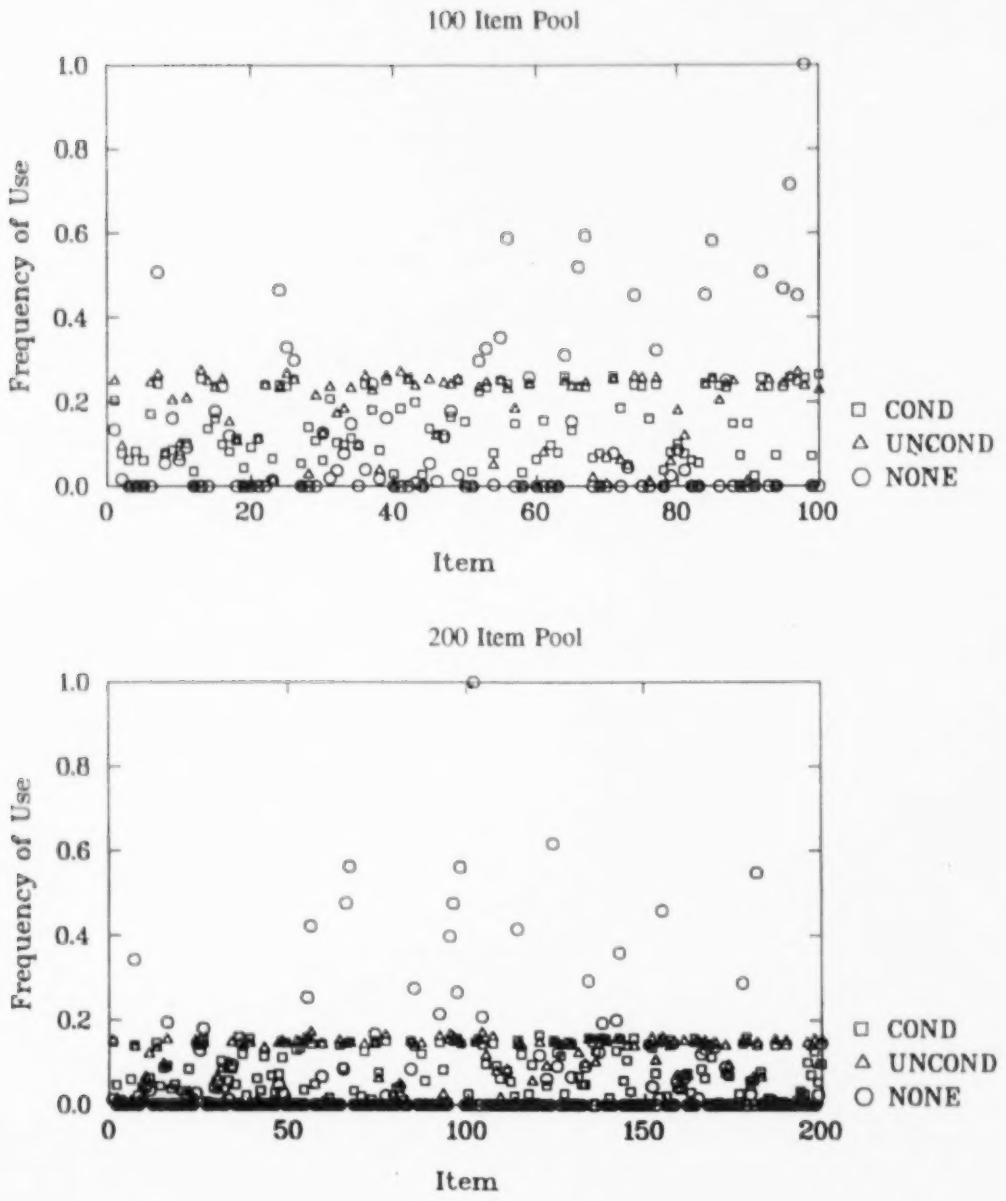


Figure 7

Average Test Information Functions